CS 7545: Machine Learning Theory **Fall 2021**

Lecture 15

Name: Tyler LaBonte **Professor:** Santosh S. Vempala

1 Boosting

Suppose $\mathcal{H}$ is a class of boolean labelings, and we have a black-box algorithm that produces a γ-weak learner, that is, $h \in \mathcal{H}$ for which

$$\Pr_{x \sim D}[h(x) = \ell(x)] \geq \frac{1}{2} + \gamma. \tag{1}$$

and $\gamma > 0$. Since a random guess would give us probability $\frac{1}{2}$, a weak learner is any learner which does better than a trivial solution. We will show that we can combine ("boost") weak learners to obtain a strong learner $h^\star$:

$$\forall \epsilon > 0, \Pr_{x \sim D}[h(x) = \ell(x)] \geq 1 - \epsilon. \tag{2}$$

A first idea is to recursively obtain weak learners for an arbitrarily large fraction of the data. That is h_1 will be correct on half the data, h_2 will be correct on half the remaining data, and so on. But this approach is too greedy, and there isn't a clear way to combine the h_i.

We will use a weighted majority to achieve a strong learner. Suppose each example $x_1, x_2, \ldots, x_n$ has weight w_i initially set to 1. When example i is misclassified, we increase w_i by a factor of $\frac{\frac{1}{2}+\gamma}{\frac{1}{2}-\gamma}$. We repeat T times using weak learners each time, then output the majority vote of the weak learners.

Suppose the number of errors of the final majority hypothesis $\hat{h}$ is $\hat{m}$. Then since at least half the weak learners got it wrong, each error has weight at least $\left(\frac{\frac{1}{2}+\gamma}{\frac{1}{2}-\gamma}\right)^{T/2}$.

Let $W(t)$ be the total weight at iteration t. Clearly $W(0) = n$. Then,

$$W(t+1) \leq \underbrace{\left(\left(\frac{1}{2} - \gamma\right)\left(\frac{\frac{1}{2}+\gamma}{\frac{1}{2}-\gamma}\right)\right)}_{\text{weighted errors}} + \underbrace{\left(\frac{1}{2} + \gamma\right)}_{\text{correct points}} W(t) = (1 + 2\gamma)W(t). \tag{3}$$

So $W(T) \leq (1+2\gamma)^T n$. Then,

$$\hat{m}\left(\frac{1+2\gamma}{1-2\gamma}\right)^{T/2} \leq (1+2\gamma)^T n. \tag{4}$$

Thus,

$$\hat{m} \leq (1 - 4\gamma^2)^{T/2} n. \tag{5}$$

Then $T = \frac{\ln n}{2\gamma^2}$ implies $\hat{m} < 1$, and since $\hat{m}$ is integer, we have $\hat{m} = 0$.

But, there is a small catch. Actually $\hat{h} \notin \mathcal{H}$, since it is a majority of concepts from $\mathcal{H}$. So to bound generalization, we need to bound the VC-dimension of the majority of k hypotheses from a class of VC-dimension d. Via the same logic as for the intersection of hypotheses, we find that this VC-dimension is $\leq 2kd \log kd$.

Theorem 1.1: Boosting

$n \geq \frac{c}{\epsilon}(Td \log(Td) \log \frac{1}{\epsilon} + \log \frac{1}{\delta}$ samples are sufficient for PAC-learning with boosting.